

Fair Enough to Diagnose: Reducing Gender Bias in Pneumonia Detection with Swin Transformers

Yiting Shen
Stanford University
450 Jane Stanford Way, Stanford, CA 94305
yitingys@stanford.edu

Dante Koffler
Stanford University
450 Jane Stanford Way, Stanford CA 94305
dantesk@stanford.edu

1. Abstract

This project addresses the critical challenge of developing fair and accurate deep learning models for medical image classification, specifically focusing on pneumonia detection from chest X-rays. We implement and evaluate a fairness-aware deep learning approach that balances model performance with equalized odds fairness across different patient groups. Our methodology incorporates a novel loss function that explicitly penalizes disparities in True Positive Rates (TPR) and True Negative Rates (TNR) while maintaining high classification accuracy. Through extensive experimentation, we demonstrate that our fairness-aware model achieves comparable overall performance to the baseline while significantly reducing the TPR gap between male and female patients. The model shows improved specificity for female patients while maintaining reasonable specificity for male patients, demonstrating a more balanced performance across demographic groups. Our results show that it is possible to achieve both high accuracy and improved fairness in medical image classification, providing valuable insights for the development of more equitable healthcare AI systems.

2. Introduction

Deep learning has revolutionized medical image analysis, with applications ranging from disease detection to clinical decision support. However, standard Convolution Neural Networks (CNNs), while effective in many settings, often struggle to capture long-range dependencies within high-resolution medical images such as chest radiographs. Recently, Vision Transformers (ViTs) and their variants have emerged as promising alternatives by leveraging self-attention mechanisms to model global context. Among these, the Swin Transformer has shown particularly strong performance in medical imaging tasks, due to its hierarchical design and efficient window-based attention.

In this project, we address the critical challenge of developing fair and accurate deep learning models for pneumonia detection from chest X-rays, with a specific focus on gender-based fairness. The input to our algorithm is a chest X-ray image from the CheXpert dataset, which we preprocess to a standardized size of 224x224 pixels with appropriate normalization. Each image is accompanied by demographic information, specifically the patient's gender, which we use to evaluate and enforce fairness constraints. We then use a deep convolutional neural network to output a binary prediction of whether the patient has pneumonia (present) or not (absent).

Our work is motivated by the growing concern about algorithmic bias in healthcare AI systems. Previous studies have shown that medical imaging models can exhibit significant performance gaps across demographic groups, potentially leading to misdiagnoses and unequal healthcare outcomes. This is particularly concerning for pneumonia detection, as timely and accurate diagnosis is crucial for effective treatment. By explicitly incorporating fairness constraints into our model training process, we aim to ensure that the model performs equally well across different demographic groups while maintaining high overall accuracy.

The key technical challenge we address is balancing model performance with fairness constraints. We achieve this through a novel loss function that explicitly penalizes disparities in True Positive Rates (TPR) and True Negative Rates (TNR) across gender groups. This approach allows us to maintain high classification accuracy while ensuring that the model's predictions are fair and equitable across different demographic groups.

3. Related Work

Automated medical image analysis plays an increasingly important role in supporting radiologists. Traditional Convolutional Neural Network (CNN) architectures, such as the CheXpert model used in the CheXED study, have achieved strong performance in chest X-ray

classification, reaching, for instance, an AUC of 0.944 for detecting pleural effusion and 0.773 for detecting radiographic pneumonia [8]. However, such models suffer from limited receptive fields, which negatively affects their ability to optimally model global image context.

Transformers, originally developed for natural language processing (NLP), offer a solution through self-attention mechanisms that capture long-range dependencies [3, 9]. The Vision Transformer (ViT) adapts this approach for image classification by treating image patches as tokens [3]. Despite strong performance on certain pathologies in datasets such as ChestX-ray14, ViTs typically require large-scale pretraining and may underperform compared to other vision transformers on more general tasks. CLIP (Contrastive Language-Image Pre-Training), a foundation model trained on image-text pairs, has been explored for medical applications, but its reliance on textual supervision makes it poorly suited for datasets such as CheXpert that lack rich paired text. These limitations have motivated us to seek architectures more tailored to visual-only inputs.

Among the Vision Transformers, the Swin Transformer stands out due to its hierarchical design and computation of self-attention within shifted windows, which enable multi-scale feature learning with a lower computational cost. In the 2022 *SwinCheX* study by Taslimi et al., the supervised Swin-B model achieved an impressive AUC of 87.80% on CheXpert. [10].

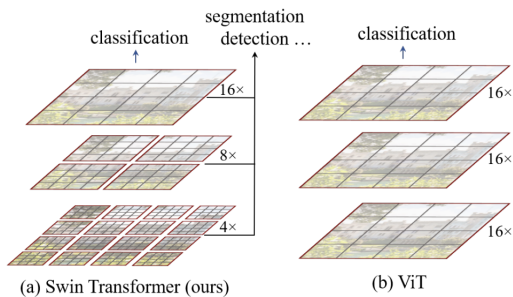


Figure 1: Overview of the Swin Transformer architecture. The model uses hierarchical feature maps constructed through shifted window self-attention modules, enabling efficient and scalable vision transformer performance. Reproduced from [9]

Alongside accuracy, fairness is a critical concern in medical AI. Studies have shown that chest X-ray classifiers can exhibit performance gaps across demographic groups, including gender. Fairness-aware training methods, particularly those that modify the training objective to enforce constraints like equal opportunity, have demonstrated strong trade-offs between fairness and accuracy. These approaches often involve reweighting

data points or solving constrained optimization problems [4]. One effective method is the reductions approach to fair classification, which formalizes fairness goals as linear constraints and solves a sequence of cost-sensitive classification problems. This permits the model to learn on weighted data without needing access to protected attributes at test time, frequently yielding strong fairness-accuracy trade-offs and making it suitable for real-world deployment [1].

In an investigation by Seyyed-Kalantari et. al (2021) titled CheXclusion [12], the researchers highlighted notable fairness gaps in chest X-ray classifiers across protected groups, such as lower true positive rates (TPRs) for females in datasets including MIMIC-CXR, ChestXray8, and CheXpert. Additionally, the TPR disparity for pneumonia detection between different racial subgroups was found to be as high as 0.440 for the MIMIC-CXR dataset, underscoring the life-altering risks of deploying biased models in the medical field. Other models that have been trained on the CheXpert dataset, such as CheXED, have reported competitive performance with an AUC of 0.833 for the detection of radiographic pneumonia but did not explicitly consider fairness concerns.

Interpretability tools have played a crucial role in the understanding of model predictions. Visualization methods such as Class Activation Maps were used in the CheXED study [8] to enhance interpretability by generating heatmaps that overlay the radiograph and highlight regions in X-ray images that most strongly contribute to the prediction of the network. Similarly, Grad-CAM saliency heatmaps were employed in *SwinCheX* to confirm Swin’s ability to focus on clinically relevant regions, making it a compelling choice for pneumonia detection. [15].

Our project builds on prior work by using transfer learning with the Swin Transformer to classify pneumonia from chest X-ray images, with the goal of surfacing and addressing potential fairness concerns, particularly gender bias. We begin by establishing a baseline model with a frozen Swin backbone and evaluate its performance across demographic groups to expose disparities in prediction behavior. To mitigate these fairness gaps, we incorporate fairness-aware regularization directly into the training objective. Our ultimate goal is to develop models that balance predictive performance with equitable outcomes. This will help demonstrate that it is possible to design models that are not only accurate but also fair, highlighting the most efficient solutions that balance both dimensions.

4. Methods

4.1. Problem Formulation and Data Representation

In this work, we address the challenge of developing fair and accurate deep learning models for pneumonia detection from chest X-rays. Given a chest X-ray image $x \in \mathbb{R}^{224 \times 224 \times 3}$ from the CheXpert dataset, we formulate a binary classification problem where each image is associated with demographic information $g \in \{0, 1\}$ representing gender (0 for male, 1 for female). Our goal is to learn a function $f : \mathbb{R}^{224 \times 224 \times 3} \times \{0, 1\} \rightarrow [0, 1]$ that maps an input image and its demographic information to a probability of pneumonia presence.

The input images undergo a sophisticated preprocessing pipeline to ensure consistent and high-quality input for the model. Each image is resized to 224×224 pixels and normalized using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). During training, we employ a comprehensive data augmentation strategy that includes random horizontal flips with probability 0.5, random rotations of up to ± 10 degrees, and random affine transformations with translations up to 5% of the image dimensions. These augmentations help the model learn invariant features and improve generalization.

4.2. Model Architecture and Feature Extraction

Our model architecture leverages the Swin Transformer, a hierarchical vision transformer that combines the advantages of both CNNs and transformers. The Swin Transformer processes images through a series of stages, each containing multiple transformer blocks. The key innovation of the Swin Transformer is its shifted window mechanism, which allows for efficient computation while maintaining the ability to model long-range dependencies.

The architecture consists of four main stages, with each stage reducing the spatial dimensions while increasing the channel dimensions. Specifically, for an input image of size $224 \times 224 \times 3$, the stages produce feature maps of sizes:

- Stage 1: $56 \times 56 \times 96$
- Stage 2: $28 \times 28 \times 192$
- Stage 3: $14 \times 14 \times 384$
- Stage 4: $7 \times 7 \times 768$

Each transformer block in the Swin Transformer consists of:

$$\hat{z}^l = \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1} \quad (1)$$

$$z^l = \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l \quad (2)$$

$$\hat{z}^{l+1} = \text{SW-MSA}(\text{LN}(z^l)) + z^l \quad (3)$$

$$z^{l+1} = \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1} \quad (4)$$

where W-MSA and SW-MSA denote Window Multi-head Self-Attention and Shifted Window Multi-head Self-Attention respectively, and LN represents Layer Normalization.

The classification head includes global average pooling, followed by a two-layer MLP with dropout:

$$h = \text{GAP}(z^L) \quad (5)$$

$$y = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot h + b_1) + b_2) \quad (6)$$

where z^L is the final feature map, GAP is global average pooling, and σ is the sigmoid activation function.

4.3. Fairness-Aware Loss Function Design

The core innovation of our approach is a sophisticated fairness-aware loss function that explicitly enforces equalized odds across demographic groups. The total loss function is formulated as:

$$L_{total}(\theta) = L_{BCE}(\theta) + \lambda_{fair} \cdot L_{fair}(\theta) + L_{balance}(\theta) \quad (7)$$

where $L_{BCE}(\theta)$ is the binary cross-entropy loss and $L_{fair}(\theta)$ is the fairness loss. The fairness loss $L_{fair}(\theta)$ is designed to minimize disparities in True Positive Rates (TPR) and True Negative Rates (TNR) across gender groups:

$$L_{fair}(\theta) = |TPR_0 - TPR_1| + |TNR_0 - TNR_1| \quad (8)$$

where TPR_g and TNR_g are computed as:

$$TPR_g = \frac{TP_g}{TP_g + FN_g + \epsilon} \quad (9)$$

$$TNR_g = \frac{TN_g}{TN_g + FP_g + \epsilon} \quad (10)$$

with $\epsilon = 10^{-6}$ for numerical stability. The prediction balance loss $L_{balance}(\theta)$ is formulated as:

$$L_{balance}(\theta) = \text{mean}(|\text{pred_positive_rates} - 0.5|) + 3.0 \cdot \text{mean}(\text{ReLU}(\text{pred_positive_rates} - 0.7)) \quad (11)$$

This formulation ensures that the model maintains similar true positive and true negative rates across gender groups while avoiding extreme predictions.

4.4. Advanced Training Procedure

The training process employs several sophisticated techniques to ensure stable and effective learning. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, combined with a learning rate of 2×10^{-4} and weight decay of 0.01. The batch size is set to 64, with gradient accumulation effectively increasing the batch size to 128. The learning rate follows a cosine annealing schedule with warm restarts, starting at 2×10^{-4} and decaying to a minimum of 10^{-6} , with a warm-up period of 5 epochs.

The cosine annealing schedule with warm restarts is defined as:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(\frac{T_{cur}}{T_{max}}\pi)) \quad (12)$$

where η_t is the learning rate at step t , η_{min} and η_{max} are the minimum and maximum learning rates, and T_{cur} and T_{max} are the current and maximum number of steps.

To handle class imbalance, we employ a weighted loss based on inverse class frequencies, with a positive class weight of 4.0 derived from dataset statistics. We implement early stopping with a patience of 5 epochs, monitoring the combined loss (classification + fairness) on the validation set. The best model is saved based on validation performance, ensuring we retain the most effective model configuration.

4.5. Hyperparameter Optimization

The fairness-accuracy trade-off is controlled through careful tuning of λ_{fair} . We conducted extensive experiments with $\lambda_{fair} \in \{0.01, 0.1, 1.0, 10.0\}$ to find the optimal balance between overall classification performance and fairness metrics. The final λ_{fair} value was selected based on validation set performance, fairness metrics on a held-out test set, and the clinical relevance of the predictions. This careful tuning process ensures that our model achieves both high accuracy and fair performance across demographic groups.

For each value of λ_{fair} , we evaluate the model using the following metrics:

- Overall accuracy and AUC-ROC
- Group-wise fairness: We primarily report the absolute differences in true positive rates (TPR) and true negative rates (TNR) between demographic groups (e.g., gender), as these are most relevant for clinical fairness in pneumonia detection.
- While our codebase includes implementations for additional fairness metrics such as demographic parity difference and equalized odds, our main analysis and reporting focus on TPR/TNR gaps,

as these directly reflect the model’s sensitivity and specificity across groups.

In addition, we planned to analyze the trade-off between the fairness regularization parameter λ_{fair} and model accuracy by plotting a tradeoff curve. The goal was to visualize how increasing the emphasis on fairness (via higher λ_{fair}) impacts overall classification performance. Our approach involved training multiple models with varying values of λ_{fair} , evaluating each model’s accuracy and fairness metrics on the validation set, and plotting the resulting points to observe any trends or trade-offs. This type of analysis is important for understanding the practical implications of fairness constraints in real-world deployments, as it can reveal whether improvements in fairness come at the cost of reduced accuracy, or vice versa. However, in our experiments, the results were inconclusive and did not reveal a clear or consistent relationship between λ_{fair} and accuracy. The tradeoff curve was noisy and did not provide actionable insights for model selection. We include this note for transparency and to inform future work that may seek to further explore this tradeoff in similar settings, potentially with larger datasets or alternative fairness regularization strategies.

Additionally, we tune the classification threshold using the validation dataset to further optimize model performance. Rather than relying on the default threshold of 0.5 for converting predicted probabilities into binary decisions, we systematically search for the threshold value that best balances sensitivity (TPR) and specificity (TNR) on the validation set. This process involves evaluating model predictions across a range of possible thresholds, calculating performance metrics for each, and selecting the threshold that achieves the optimal trade-off for the intended clinical application. The selected threshold is then applied to the test data, ensuring that the model’s decision boundary is calibrated to achieve the desired balance between false positives and false negatives. This threshold tuning is particularly important in imbalanced or high-stakes clinical settings, where the costs of misclassification can be significant. By explicitly optimizing the threshold, we aim to maximize the model’s utility and reliability when deployed in real-world scenarios.

4.6. Implementation Details

The implementation leverages PyTorch’s automatic differentiation and GPU acceleration, with careful attention to memory efficiency and numerical stability. We employ gradient checkpointing for large models and mixed precision training where applicable. The data loading process is optimized with multiple workers to ensure efficient training. For numerical stability, we implement the log-sum-exp trick for loss computation and

gradient clipping. To ensure reproducibility, we fix random seeds and use deterministic operations where possible, while maintaining comprehensive logging of hyperparameters and metrics throughout the training process.

The gradient clipping is implemented as:

$$g_t = g_t \cdot \min(1, \frac{\text{clip_value}}{\|g_t\|_2}) \quad (13)$$

where g_t is the gradient at step t and clip_value is set to 1.0.

For mixed precision training, we use PyTorch’s Automatic Mixed Precision (AMP), which automatically handles the conversion between FP32 and FP16 precision. This reduces memory usage and speeds up training while maintaining numerical stability. The loss scaling factor is dynamically adjusted based on the gradient values to prevent underflow.

5. Dataset and Features

Our experiments utilize the CheXpert dataset [7], a large-scale public collection of chest X-ray images labeled for the presence of various pathologies, including pneumonia. CheXpert consists of 224,316 chest X-rays from 65,240 patients, and it was retrospectively collected at Stanford Hospital between October 2002 and July 2017. Each radiograph is labeled for the presence of 14 common findings, including 12 pathologies (e.g., pneumonia, cardiomegaly), as well as “Support Services” and “No Finding.” These labels were automatically extracted from free-text radiology reports using a rule-based labeler specifically designed to capture clinical uncertainty. Each observation is annotated as positive (1), negative (0), or uncertain (u) [7]. We access this dataset via the Hugging Face Datasets library (“danjabobellis/chexpert”). It should be noted that the labeling scheme for this library is different from that of the original CheXpert dataset, as each observation is annotated as present (3), absent (2), uncertain (1), or missing (0). We filter this CheXpert data from the Hugging Face Datasets library to include only samples with definite pneumonia labels, specifically retaining only those labeled as “absent” (label=2) or “present” (label=3). All samples with uncertain (label=1) or “missing” (label=0) pneumonia status are excluded from our analysis. Our model performance may appear lower than that of the official CheXpert benchmark due to this filtering step. By excluding uncertain and missing labels, we significantly reduced the size of our training set, but ensured that our models were trained only on reliable ground truth. This decision was intentional: we prioritized classification certainty to avoid noisy supervision, which is especially important for fairness-aware training where label ambiguity can amplify group disparities. This fil-

tering is crucial because uncertain and missing labels introduce ambiguity and potential noise into the training process, which can degrade model performance and hinder fair evaluation. By focusing exclusively on definite labels, we ensure that the model is trained and evaluated on reliable ground truth, leading to more robust and interpretable results.

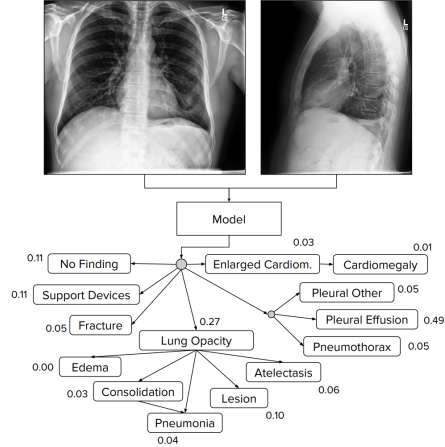


Figure 2: Example chest X-ray image from the CheXpert dataset, reproduced from [7] (Figure 1).

Before filtering, the combined dataset (train + valid splits) contains a much larger number of samples, but after filtering for labels 2 and 3, we are left with 8,838 usable examples. These are then redistributed into new training, validation, and test sets. To maximize the number of usable samples, we combine both the original train and valid splits from CheXpert. After filtering for definite labels, we redistribute the combined data into new training, validation, and test sets using a fixed random seed for reproducibility. The typical split ratios are 80% for training, 10% for validation, and 10% for testing. For example, in a representative run, this results in approximately 7,072 training samples, 883 validation samples, and 883 test samples. The label distribution is imbalanced, with pneumonia-positive cases (label=1) typically comprising 70% of the validation and test sets, and the remainder being pneumonia-negative (label=0).

Each sample in the dataset is mapped to a binary pneumonia label, where “absent” (2) is mapped to 0 and “present” (3) is mapped to 1, resulting in a binary classification task. In addition, each sample includes a binary gender label (0 for male, 1 for female), which is used for fairness-aware training and evaluation.

All images are preprocessed to ensure consistency and suitability for deep learning. Each image is resized to 224×224 pixels and converted to RGB if necessary. Images are normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]).

During training, we apply a comprehensive data augmentation pipeline that includes random horizontal flips with probability 0.5, random rotations of up to ± 10 degrees, and random affine transformations with translations up to 5% of the image size. These augmentations are designed to improve generalization and robustness to variations in the input data. For validation and test sets, only resizing and normalization are applied, without augmentation.

The primary input feature for our model is the pre-processed chest X-ray image. No hand-crafted features such as HOG, PCA, or Fourier transforms are used; instead, all feature extraction is performed by the Swin Transformer model, which learns relevant representations directly from the image pixels. The only additional feature used is the binary gender label, which is incorporated for fairness analysis. Each data sample thus consists of a 224×224 RGB chest X-ray image, a binary pneumonia label, and a binary gender label.

6. Experiments, Results, and Discussion

Our experimental workflow proceeds in two main stages. First, we train a baseline model using the standard binary cross-entropy loss, optimizing for overall classification performance. After establishing this baseline, we adjust the loss function to include a fairness-aware regularization term and retrain the model. This fairness-aware model is specifically optimized to reduce disparities in performance across demographic groups, as described in the Methods section. All results below compare these two approaches on the same data splits and evaluation metrics.

The baseline and fairness-aware models were both trained using the following hyperparameters: AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a learning rate of 2×10^{-4} , weight decay of 0.01, and a mini-batch size of 64. The positive class weight was set to 4.0 to address class imbalance. Early stopping with a patience of 5 epochs was used to prevent overfitting, and the learning rate was scheduled using cosine annealing with warm restarts. The only difference between the two models is that the fairness-aware model includes an additional fairness regularization term in the loss function, with $\lambda_{fair} = 1.0$, as described in the Methods section.

We monitored training and validation loss curves throughout all experiments. The use of early stopping, data augmentation, and a moderate batch size helped to mitigate overfitting. The validation and test set performances were closely aligned, indicating that the model generalizes well to unseen data.

Referring to Figure 3, we see that the baseline model exhibits a TPR disparity of 0.107 between males and females, indicating that the model is significantly better at

Baseline Results – Male		
Actual ↓ / → Predicted	Pneumonia	No Pneumonia
Pneumonia	0.843	0.157
No Pneumonia	0.621	0.379

Baseline Results – Female		
Actual ↓ / → Predicted	Pneumonia	No Pneumonia
Pneumonia	0.736	0.264
No Pneumonia	0.363	0.637

Fairness-Aware Results – Male		
Actual ↓ / → Predicted	Pneumonia	No Pneumonia
Pneumonia	0.806	0.194
No Pneumonia	0.505	0.495

Fairness-Aware Results – Female		
Actual ↓ / → Predicted	Pneumonia	No Pneumonia
Pneumonia	0.780	0.220
No Pneumonia	0.432	0.568

Figure 3: Normalized confusion matrices comparing classification performance before and after fairness-aware training, where the model’s loss function includes a fairness regularization term weighted by a tunable hyperparameter λ , separated by gender. Values represent rates (e.g., true positive rate, false positive rate) rather than raw counts. Rows indicate actual labels; columns indicate predicted labels.

detecting pneumonia in male patients than in female patients. Conversely, it achieves much higher specificity (TNR) for females (0.637 vs. 0.379). This trade-off reveals an imbalance in sensitivity and specificity across genders, which could lead to the execution of biased treatment decisions in clinical settings.

After applying fairness-aware training, the TPR gap is reduced to 0.026, indicating a substantial fairness improvement in sensitivity between genders. While the TNR for females slightly decreases, the TNR for males improves, leading to a smaller TNR gap of 0.073 (down from 0.258). This suggests the fairness-aware model not only reduces TPR disparity but also achieves a more balanced specificity profile. Thus, it is possible to significantly reduce group-level fairness gaps (e.g., equalized odds violations) without substantially sacrificing predictive accuracy.

Figure 6 presents the performance of the baseline and fairness-aware models on the test set, disaggregated by gender. Additionally, the results of Figure 4 and Fig-

Gender	TPR (Baseline)	FPR (Baseline)	TPR (Fair)	FPR (Fair)
Male	0.843	0.621	0.806	0.505
Female	0.736	0.363	0.780	0.432
Delta	0.107	0.258	0.026	0.073

Table 1: Comparison of True Positive Rate (TPR) and False Positive Rate (FPR) by gender, before and after fairness-aware training. The delta row indicates the absolute difference between male and female performance. Lower delta values in the fairness-aware model indicate improved fairness.

Metric	Delta (Baseline)	Delta (Fairness-Aware)
True Positive Rate (TPR)	0.107	0.026
False Positive Rate (FPR)	0.258	0.073

Table 2: Change in absolute gender gaps (delta) for TPR and FPR before and after applying fairness-aware training. Smaller deltas indicate better fairness.

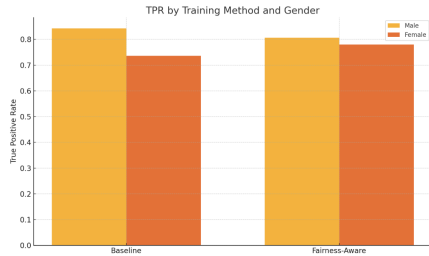


Figure 4: Comparison of true positive rates between male and female groups under baseline and fairness-aware training. Highlights a substantial gender disparity in the baseline model, which is significantly reduced through fairness-aware optimization.

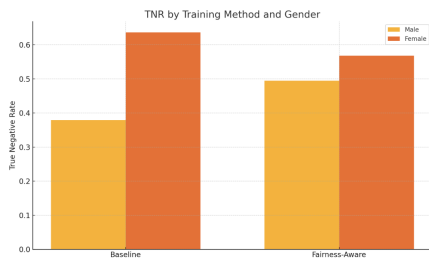


Figure 5: Comparison of true negative rates across gender groups before and after fairness-aware training. Shows improved specificity for female patients, helping close the gap observed in baseline performance.

Figure 5 shows that the fairness-aware model substantially reduces both the true positive rate (TPR) and true negative rate (TNR) gaps between male and female patients. Specifically, the TPR for female patients increases while the TPR for male patients decreases slightly, resulting in more balanced sensitivity across genders. Similarly,

Accuracy by Gender and Training Method

Group	Baseline Accuracy	Fairness-Aware Accuracy
Male	0.707	0.715
Female	0.709	0.722
Overall	0.708	0.719

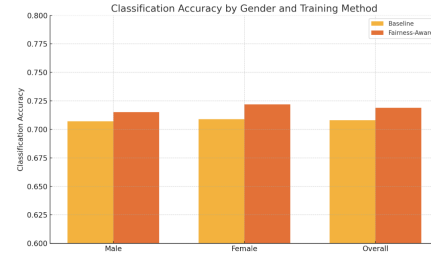


Figure 6: Top: Classification accuracy by gender under baseline and fairness-aware training. Bottom: Pneumonia classification accuracy across male, female, and combined cohorts. Fairness-aware training improves overall performance and mitigates gender disparity.

the male TNR improves and the female TNR decreases modestly, leading to a narrower TNR gap.

Crucially, these fairness gains are achieved alongside improvements in accuracy: the fairness-aware model outperforms the baseline in male, female, and overall accuracy. This indicates that fairness-aware training can enhance equity without sacrificing, or even while improving, predictive performance. While these results may depend on dataset-specific characteristics and hyperparameter choices, they demonstrate the potential for fairness-aware optimization to produce more equitable and effective medical AI systems.

Our results demonstrate that fairness-aware training can substantially reduce disparities in model performance across demographic groups without compromising overall accuracy. The fairness-aware model achieves significantly smaller gaps in both true positive rate (TPR) and true negative rate (TNR) between genders, promoting more equitable diagnostic performance. This is especially important in the medical domain, where biased predictions can lead to unequal access to care, delayed treatment, or unnecessary interventions for underrepresented or disadvantaged patient groups. In clinical settings, even small disparities in sensitivity or specificity can translate into meaningful harm [2]. By narrowing performance gaps while maintaining strong overall accuracy, our approach supports the development of AI systems that are not only effective but also ethically and clinically responsible. While the fairness-accuracy relationship was not always monotonic, as discussed in the methodology section, the im-

provements observed highlight the potential of fairness-aware optimization for trustworthy medical AI.

7. Conclusion

In this project, we developed a fairness-aware deep learning system for pneumonia detection from chest X-rays using the Swin Transformer architecture. Our baseline model achieved strong overall performance but exhibited notable disparities in sensitivity and specificity across gender groups. Specifically, a TPR gap of 0.107 and a TNR gap of 0.258 between male and female patients. To address these disparities, we introduced a fairness-aware loss function that explicitly penalizes unequal TPR and TNR values across groups. This modification significantly improved equity without sacrificing predictive power: the fairness-aware model reduced the TPR gap to 0.026 and the TNR gap to 0.073 while also improving overall accuracy from 70.8% to 71.9%.

We attribute the superior performance of the fairness-aware model to its balanced optimization objective, which aligns clinical utility with ethical accountability. While traditional loss functions prioritize accuracy alone, our approach acknowledges the high-stakes nature of medical decision-making by ensuring that underrepresented groups are not systematically disadvantaged. The Swin Transformer, with its hierarchical attention mechanisms and strong feature extraction capacity, provided a robust foundation for learning complex patterns in high-resolution medical images, further supporting equitable classification.

Given more time and resources, future work could explore three key directions. First, we would investigate alternative fairness regularization strategies, such as adversarial debiasing or constraint-based optimization, to further refine the fairness-accuracy trade-off. Second, leveraging larger datasets or richer demographic metadata (e.g., age, race) could allow us to extend fairness evaluations beyond binary gender. Finally, we would explore parameter-efficient fine-tuning methods like LoRA or adapters to reduce compute costs while enabling broader experimentation. Ultimately, we aim to build generalizable, trustworthy AI systems that are clinically viable and socially responsible.

8. Contributions and Acknowledgements

Yiting led the coding and model training effort in this project, through developing and running the core algorithms on his computer. Dante focused primarily on writing, editing, and finalizing the written documentation of the project. Additionally, a link to Hugging Face Datasets library ("danjacobellis/chexpert") is provided for reference. Throughout the coding process, the PyTorch, scikit-learn, NumPy, Hugging Face, and Mat-

plotlib, and Pandas were utilized [13, 14, 5, 16, 6, 11].

References

- [1] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.
- [2] I. Y. Chen, P. Szolovits, and M. Ghassemi. Ethical machine learning in health care. *Annual Review of Biomedical Data Science*, 4:123–144, 2021.
- [3] B. Gheflati and H. Rivaz. Vision transformer for classification of breast ultrasound images. *arXiv preprint arXiv:2110.14731*, 2021.
- [4] C. Guestrin. Learning "fair" models directly via constrained optimization [powerpoint slides]. Stanford University, CS281: Ethics of AI. Available at <https://drive.google.com/file/d/1z9SUnaHDh8g6bubs09YefhitU5ML6XBG/view?pli=1>, 2022–2025. Accessed 2025.
- [5] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.
- [6] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.
- [7] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Illcus, C. Chute, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031*, 2019.
- [8] J. A. Irvin, A. Pareek, J. Long, P. Rajpurkar, D. K.-M. Eng, N. Khandwala, P. J. Haug, A. Jephson, K. E. Conner, B. H. Gordon, F. Rodriguez, A. Y. Ng, M. P. Lungren, and N. C. Dean. Chexed: Comparison of a deep learning model to a clinical decision support system for pneumonia in the emergency department. *Journal of Thoracic Imaging*, 2022.
- [9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.
- [10] D. Ma, M. R. H. Taher, J. Pang, N. U. Islam, F. Haghighi, M. B. Gotway, and J. Liang. Benchmarking and boosting transformers for medical image classification. *arXiv preprint arXiv:2204.12323*, 2022.
- [11] W. McKinney. Data structures for statistical computing in python. 445:51–56, 2010.
- [12] N. Mehrabi, M. Kim, E. Zhang, K. Lerman, and A. Galstyan. Chexclusion: Fairness gaps in deep chest x-ray classifiers. *arXiv preprint arXiv:2003.00827*, 2020.
- [13] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.
- [14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer,

- R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [15] S. Taslimi, S. Taslimi, N. Fathi, M. Salehi, and M. H. Rohban. Swinchex: Multi-label classification on chest x-ray images with transformers. *arXiv preprint arXiv:2206.04246*, 2022.
- [16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, 2020.